Activity modelling using journey pairing of taxi trajectory data

Shuhui Gong*†, John Cartlidge‡, Ruibin Bai§, Yang Yue† Qingquan Li† and Guoping Qiu¶
*International Doctoral Innovation Centre, University of Nottingham Ningbo China, Ningbo China
†Department of Urban Informatics, Shenzhen University, Shenzhen China
‡Department of Computer Science, University of Bristol, Bristol UK
§Department of Computer Science, University of Nottingham Ningbo China, Ningbo China
¶Department of Computer Science, University of Nottingham, Nottingham UK
Email: shuhui.gong@nottingham.edu.cn, john.cartlidge@bristol.ac.uk, ruibin.bai@nottingham.edu.cn
yueyang@szu.edu.cn, liqq@szu.edu.cn, guoping.qiu@nottingham.ac.uk

***Abstract*—Taxi GPS data offers an opportunity to discover behavioural patterns in urban populations. However, the raw data does not provide a link between outbound and return journeys of individual travellers. Without this information, it is not possible to track individual behaviours. In this study, we propose a method for pairing taxi journeys and apply it to taxi trajectory data for the city of Shenzhen, China. Journeys related to three activities are considered: shopping, medical, and work. Results, validated using questionnaire data collected in Shenzhen, reveal behavioural patterns and suggest possibilities for applications in urban design.**

***Index Terms*—Power law distance decay function; Monte Carlo simulation; travel behaviour analysis;**

I. Introduction

GPS data has been widely used on travel behaviour analysis, such as geographical model calibration [1], discovering travel patterns [2], and modelling demand for points of interest (POI) [3]. However, it is challenging to infer behavioural activities of individuals from taxi trajectory data alone. Although GPS taxi data includes accurate individual locations, how to discover the correlations between journeys and trip purposes remains a difficult and unsolved technical challenge. The location data is rich, but the activity information is sparse [2], [4].

Here, we propose a preliminary "paired journey model" to estimate return journeys in taxi trajectory data, and discover the relationship between predecessor and successor activities. In particular, we analyse three activities: shopping related, medical related, and work related. This work follows previous studies using social demographics [5], time series prediction [6], and social media reviews [7] to enrich taxi trajectory data with the aim of building behavioural models of urban activity.

The paper is organised as follows: Section II reviews related work; Section III describes the data; Section IV provides a detailed description of the methodology used; Section V presents preliminary results; and finally, Section VII concludes.

II. Related Work

Previously, some research show that the sequence of activities determines the mobility patterns, and there is a relationship between predecessor activity and the successor activity [8] [9]. However, the activity-based analysis currently is conducted through travel diary datasets, which is expensive and time consuming to gather, and therefore often small in scale. To address this, we develop a paired journey model using taxi data to automatically extract return journeys, and show the relationship between predecessor activity and successor activity. Moreover, we also explore the probability that a person will return directly to their origin after a predecessor activity.

Distance decay function is first proposed and calibrated in 1981 [10]. It is now widely used on estimating trip patterns from GPS data [11], and inferring trip purpose [2]. Liu extends the distance decay effect into power law distance decay function by considering the power value as a parameter and varying in different situations, which has been proved to have good performance on travel behaviour analysis [12]. The expression is:

$$Pr(O_i|(x,y)) = Pr((x,y)|O_i) \propto A_i d((x,y),O_i)^{-\beta} \quad (1)$$

where $Pr(O_i|(x,y),t)$ represents the probability that a journey visit is intended for POI activity O_i, A_i is a constant, and $d((x,y),O_i)^{-\beta}$ is the distance between customers location (x,y) to POI location O_i. In previous research β is optimized as 1.5 [13], [14], and [15]. Therefore in this study, we use power law distance decay function and directly use 1.5 as beta value to build a paired journeys model.

III. Data

We use GPS trajectory data for more than ten million taxi journeys in Shenzhen, China, between 24 September 2015 and 20 October 2015. Each journey includes pick-up time, drop-off time, pick-up location, drop-off location, and date. For model validation, we use 712 questionnaires about people's behavioural habits collected in Shenzhen's major shopping areas. Two questions in the survey are pertinent to this study: (Q1) How long did you travel to the shopping area? Options: less than 10 minutes, 10-20 minutes, 20-30 minutes, and more than 30 minutes. (Q2) How long do you intend to stay in the shopping area? Options: below 1 hour, 1-2 hours, 2-4 hours, and over 4 hours.

Algorithm 1 Monte Carlo identification of activity purpose

Input: a set of filtered journeys
Output: journey purpose

for each drop off point **do**
  calculate the probabilities of n journeys as return journeys to activities $(p_1, p_2, \ldots, p_n)$ ▷ $\sum_{i}^{n} p_i = 1$
  set $p_0 = 0$
  generate random value, $r \in [0, 1]$
  for $i = 0$ to $n - 1$ **do** ▷ decide return journeys
    if $\sum_{j=0}^{i} p_j \leq r < \sum_{j=0}^{i+1} p_j$ **then**
      result = journey$[i + 1]$
    end if
  end for
end for
return result

IV. Methodology

The process to build the paired journey model is shown in Fig. 1. To simplify the problem, we select isolated POIs and assume that taxi journeys with drop-off points (DOP) close to that POI are aiming for that POI. In this study, we select three POIs with different activities: a large IKEA store (a shopping POI), a large hospital (the Third Hospital; a medical related POI), and a large office building (Tencent; a work related POI). To discover return journeys, we select all taxi journeys from Sep 24 to Oct 20 in 2015. Three rules are used to identify return journeys: (i) out and return journeys are in same day. For medical related journeys, we do not consider the situation that patients live in the hospital. (ii) The difference between outward journey drop off time and return journey pick up time must be positive. (iii) d_1 represents Euclidean distance from DOP in outbound journey to PUP in return journey; d_2 represents distance from PUP in outbound journey to DOP in return journey. When d_1 and d_2 are small, it is more likely that the two journeys are a return journey "pair". The challenge is how to determine the distance that people walk to their intended destination after alighting their taxi (i.e., suitable values of d_1 and d_2). Here, we explore pairing return journeys by increasing d from 0 to 500m in steps of 10m (using the 500m upper bound of walking distance from taxi to destination presented by [1]). Only return journeys that satisfy all three rules are considered as a possible return pairing.

After discovering all possible return journeys, we next consider the number of possible return journeys, J_{ret}, to each outbound journey: if $J_{ret} = 0$, we consider the travel journey has no direct return journey (i.e., outward journey is non-paired); if $J_{ret} = 1$, we consider the one possible match as the return journey; if $J_{ret} > 1$, Monte Carlo simulation (refer to Algorithm 1) is used to select the one most likely return journey based on d_1 and d_2 (journeys with smaller d_1 and d_2 have more chance of selection). Once journey pairing is complete, we discover and analyse the travel patterns for the three activity types, and evaluate the accuracy of the paired journey model using the questionnaire data collected in Shenzhen.

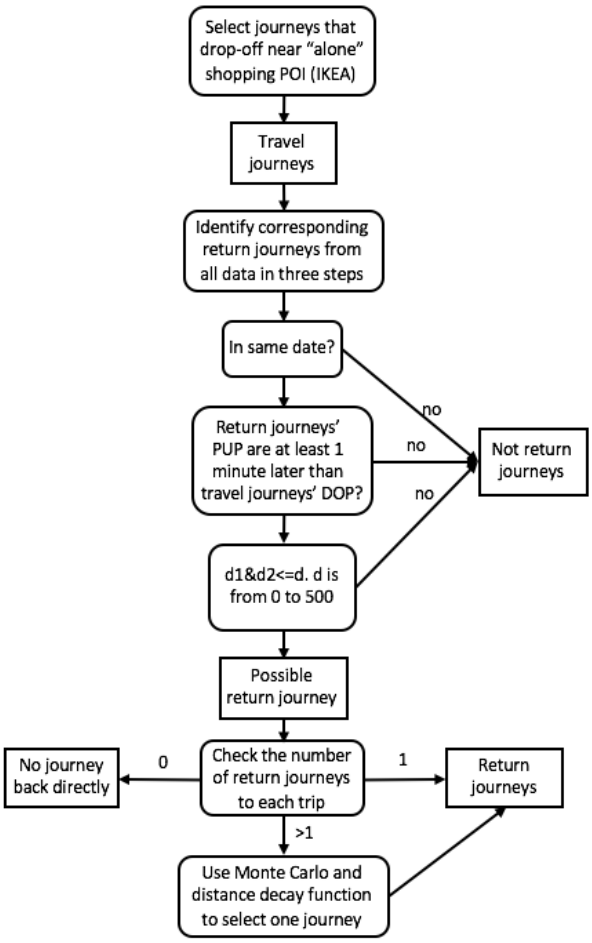


Fig. 1: Framework to select return journeys: d_1 is Euclidean distance from DOP in outward journey to PUP in return journey; d_2 is distance from PUP in outward journey to DOP in return journey.

V. Result and discussion

We collect 3,075 journeys with DOP near IKEA; 4,103 journeys with DOP near Third hospital, and 1,048 journeys with DOP near Tencent building. We discover that $d_2 = 250$ best pairs journeys. The distance sample from $d_2 = 100$, $d_2 = 200$, and $d_2 = 250$ are shown in Fig. 2. It is because: (i) when $d_2 = 100$, the resolution of the distance between the out PUP and return DOP is approximately the distance to cross a road. Therefore, we infer that when $d_2 \leq 100$, the journeys are a return pair; (ii) when $d_2 = 200$, the distance is similar to half of the width of a residential estate; (iii) $d_2 = 250$ is similar to a distance from one gate to another in a residential estate (for example, from south gate to north gate). When $d_1 > 250$, we find some situations where the two points are not located near one POI. Therefore, we filter for possible return journeys, J_{ret}, such that $d_1 \leq 250$ and $d_2 \leq 250$. The results show that 55% of shopping journeys have a return taxi trip back to origin; 61% of medical journeys have a return taxi trip back to origin; and 62% of work journeys have a return taxi trip

(a) 100 m (b) 200 m (c) 250 m

Fig. 2: Distance sample (d2) of 'paired' journeys. Blue dot: outbound PUP. Green dot: return DOP.

back to origin.

Fig. 3 shows the travel time distributions for the three activities. We see that journeys with short travel time are more likely to be paired. In particular, 64% of shopping journeys with travel time within 11 minutes will travel back to origin after shopping, 70% of medical-related journeys with travel time within 17 minutes will return to origin, and 71% of work journeys with travel time within 14 minutes will return to origin. For other travel times, the proportion of whether people return to origin location is roughly 50% (ratio between red (paired) and black (non-paired) lines). From the results, we infer that people are more likely to return to their original places after a short trip.

Fig. 4 shows the drop-off time distribution for the three activities. 60% of people will return to the origin location if they go shopping between 10am and 3pm; 70% of people return to the origin location if they visit a hospital between 6 am and 11 am. 63% of people return to the origin location when they go to work between 8 am and 11am, and 70% of them return to the origin if they go to work between 1 pm and 3 pm. At other times, the proportion is approximately 50%. From the results, we infer that people who go to see the doctor in the morning, or go shopping at noon, or go to work in the afternoon are much more likely to go straight back to their original place after predecessor activities.

Fig. 5 shows the time that people spend on different activities (in minutes). We see that people usually spend much more time (up to 14 hours) on medical treatment and work, compared with shopping (8 hours maximum). In particular, 81% of shopping activities last less than 4 hours, while 46.7% of hospital activities last more than 4 hours. Moreover, people tend to work for 6 to 7 hours, but they are more likely to spend a shorter time (less than two hours) for shopping and medical activities. This difference is what we would intuitively expect.

Fig. 6 shows the destinations of journeys returning after each activity. We see that 65% of people travel back to residential locations after shopping, 70% return to residence after medical activities, while only 48% of people travel back home after work. We therefore infer that people are more likely to go shopping after work (25%) than after shopping (6.2%) and medical service (5.5%). We also see that while there are relatively few journeys aimed for entertainment after medical activities (0.5%; compared with 7.7% after shopping activities), the proportion of journeys which aim for another hospital after medical activity (6.7%) is higher than after shopping (only 1.4%) and after work (5%). It is interesting, because it indicates that patients travel between hospitals after each visit. This could be because hospitals offer different specialisms.

VI. Evaluation of Shopping and Work Activities

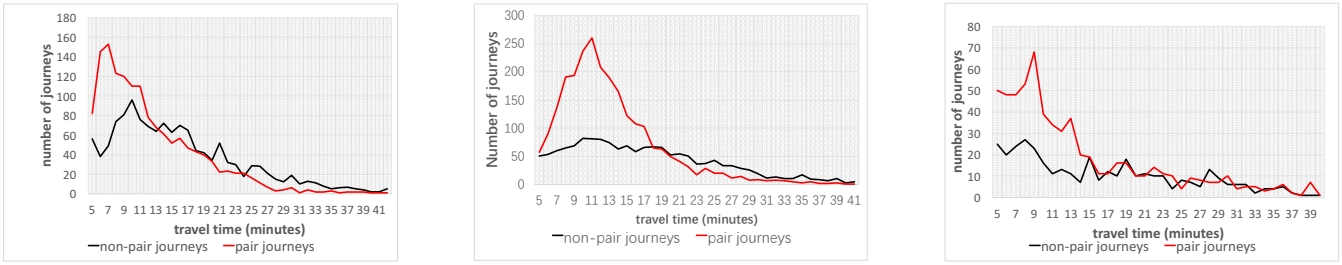
Here, we use survey data in Shenzhen about shopping behaviours as ground truth to test the performance of IKEA pairing shopping journeys. Two dimensions are used: (i) travel time; and (ii) time spent on shopping.

Fig. 7 presents validation results on travel time distribution and time spent distribution on shopping. It is clear that IKEA pairing journeys has similar distribution to surveys. Meanwhile, we also use Mean Absolute Percentage Error as criteria to test the performance of IKEA shopping journeys, which is only 4.38% on travel time, and 3.27% on shopping time evaluation. Therefore, the paired journey model has high performance with low percentage error.

We also use ground truth in data from [9] about work-related journeys (obtained from agent-based simulation) to test the drop-off time distribution extracted from taxi journeys to Tencent, which is shown in Fig. 8. We see that the results curve is similar to observation journey proportions. From the validation results, we see that the paired journey model has a good performance on estimating whether people will travel back to their original locations after predecessor activities.

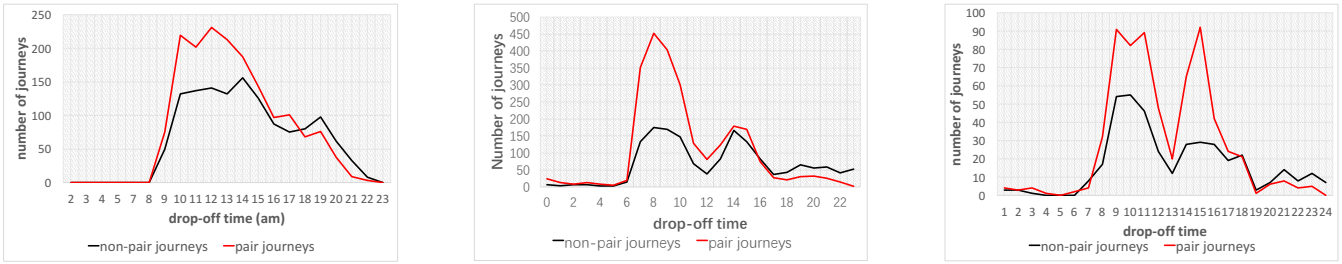
VII. Conclusion

In this paper, we build a paired journey model to infer people's trips after shopping, taking medical treatment, or working. Results demonstrate that the paired journey model has a good performance on extracting outward and return journeys. The results also demonstrate that people are more likely to return directly to their starting location after a short trip. In particular, people who go to see the doctor in the



(a) shopping journeys (b) medical journeys (c) work journeys

Fig. 3: Travel time distribution. Black line plots journeys with no return pairing; red line plots paired journeys.



(a) shopping journeys (b) medical journeys (c) work journeys

Fig. 4: Drop-off time distribution. Black line plots journeys with no return pairing; red line plots paired journeys.

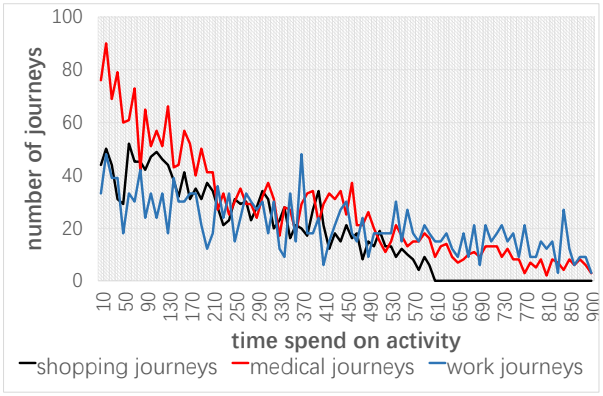


Fig. 5: Distribution of time spent on activities.

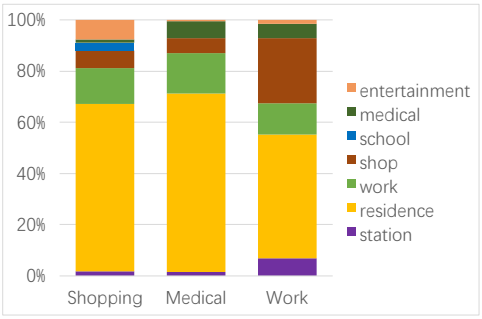
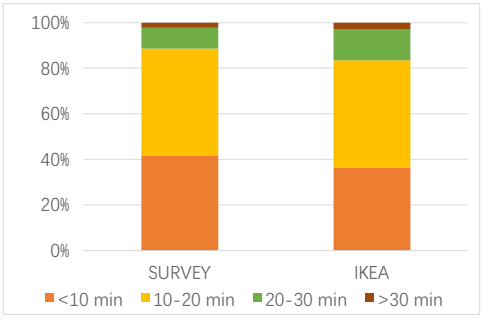


Fig. 6: Paired journeys: relationship between predecessor and successor activities.

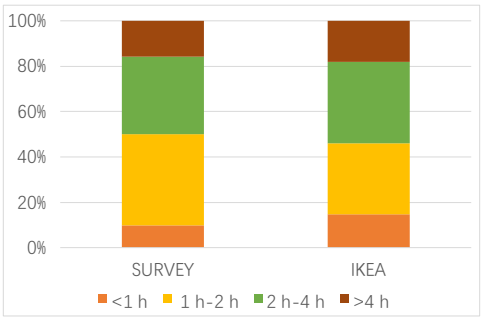
morning, or go shopping at noon, or go to work in the afternoon are much more likely to return directly back to their starting location after predecessor activities.

We also see that a large proportion of people travel from their residential locations and return home after shopping or hospital visits, while one quarter of people prefer to go shopping after they end work. Moreover, after one hospital visit, roughly 7% of journeys head to another hospital, which is an interesting finding to see some patients move to another medical institution after taking medical treatment.

The paired journey model has multiple potential applications: (1) It could be used to predict people's successor activities based on predecessor activities, which could be contributed to understand human's daily movements. Companies can also apply it to understand the customer's daily routine, and do advertisement for target customers. (2) In medical field, since some people will go to different hospitals after taking medical care, the paired journey model could use



(a) Travel times distribution



(b) Time spent shopping

Fig. 7: Pair journeys process validation using questionnaire data based on two dimensions: travel times, and time spend on shopping. Three results are compared: trip diaries in survey, IKEA journeys, and shopping journeys discovered from AIM.

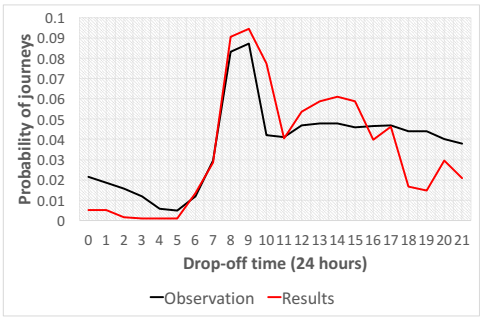


Fig. 8: Validation of drop-off time distribution for work. The observations are ground truth in data taken from [9], and the results are taxi journeys discovered in this study.

individual travel information in GPS data to infer whether they are satisfied about the treatment (by discovering return activities). The results could be directly applied to medical online platforms, which provide pre-examination to patients.

Acknowledgment

This research was supported by Zhejiang Natural Science Foundation (Grant No. LR17G010001), Ningbo Science and Technology Bureau (Grant No. 2014A35006), UK Engineering and Physical Sciences Research Council (Grant No. EP/L015463/1), Refinitiv (formerly Thomson Reuters Financial and Risk), the National Science Foundation of China (No. 41671387, 91546106, 71471092), Shenzhen Scientific Research and Development Funding Program (No. CXZZS20150504141623042).

References

[1] Y. Yue, H. Wang, B. Hu, Q. Li, Y. Li, and A. Yeh, "Exploratory calibration of a spatial interaction model using taxi GPS trajectories," *Computers, Environment and Urban Systems*, vol. 36, no. 2, pp. 140–153, 2012.

[2] L. Gong, X. Liu, L. Wu, and Y. Liu, "Inferring trip purposes and uncovering travel patterns from taxi trajectory data," *Cartography and Geographic Information Science*, vol. 43, no. 2, pp. 103–114, 2016.

[3] Y. Liu, C. Liu, X. Lu, H. Zhu, H. Zhu, and H. Xiong, "Point-of-interest demand modeling with human mobility patterns," in *ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2017, pp. 947–955.

[4] P. Wang, Y. Fu, G. Liu, W. Hu, and C. Aggarwal, "Human mobility synchronization and trip purpose detection with mixture of hawkes processes," in *Proc. 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2017, pp. 495–503.

[5] S. Gong, J. Cartlidge, Y. Yue, G. Qiu, Q. Li, and J. Xin, "Geographical huff model calibration using taxi trajectory data," in *Proc. 10th ACM SIGSPATIAL Workshop on Computational Transportation Science*, 2017, pp. 30–35.

[6] J. Cartlidge, S. Gong, R. Bai, Y. Yue, G. Qiu, and Q. Li, "Spatio-temporal prediction of shopping behaviours using taxi trajectory data," in *Proc. 3rd IEEE Int. Conf. on Big Data Analysis*, 2018, pp. 112–116.

[7] S. Gong, J. Cartlidge, R. Bai, Y. Yue, G. Qiu, and Q. Li, "Automated prediction of shopping behaviours using taxi trajectory data and social media reviews," in *Proc. 3rd IEEE Int. Conf. on Big Data Analysis*, 2018, pp. 117–121.

[8] C. Bhat and F. Koppelman, "Activity based modeling of travel demand, handbook of transportation science, editor: Rw hall," 2003.

[9] L. Wu, Y. Zhi, Z. Sui, and Y. Liu, "Intra-urban human mobility and activity transition: Evidence from social media check-in data," *PloS one*, vol. 9, no. 5, p. e97010, 2014.

[10] A. S. Fotheringham, "Spatial structure and distance-decay parameters," *Annals of the Association of American Geographers*, vol. 71, no. 3, pp. 425–436, 1981.

[11] Y. Liu, C. Kang, S. Gao, Y. Xiao, and Y. Tian, "Understanding intra-urban trip patterns from taxi trajectory data," *Journal of geographical systems*, vol. 14, no. 4, pp. 463–483, 2012.

[12] Y. Liu, L. Gong, and Q. Tong, "Quantifying the distance effect in spatial interactions," *Acta Scientiarum Naturalium Universitatis Pekinensis*, vol. 50, no. 3, pp. 526–534, 2014.

[13] Y. Li, S. Gong, and H. Liddell, "Support vector regression and classification based multi-view face detection and recognition," in *Proc. 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2000, pp. 300–305.

[14] S. Gao, Y. Wang, Y. Gao, and Y. Liu, "Understanding urban traffic-flow characteristics: a rethinking of betweenness centrality," *Environment and Planning B: Planning and Design*, vol. 40, no. 1, pp. 135–153, 2013.

[15] C. Kang, X. Ma, D. Tong, and Y. Liu, "Intra-urban human mobility patterns: An urban morphology perspective," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 4, pp. 1702–1717, 2012.